**مقدمه**

**ایده ها و روش های پیشنهادی**

**معرفی استخراج دانش[[1]](#footnote-1) و شناسایی موجودیت های نامدار[[2]](#footnote-2)**

استخراج دانش و شناسایی موجودیت های نامدار دو روش برای بدست آوردن دانش ساخت یافته ساده از متن هستند. هدف از سیستم های استخراج دانش یافتن و درک بخش های مرتبط بهم در متن به صورت محدود است. بسیاری از این سیستم ها، به استخراج دانش از درون تکه های متن می­پردازند و یک نمایش ساختیافته از اطلاعات مرتبط بهم در متن را ارائه می­کنند. از اطلاعات بدست آمده در این سیستمها می­توان برای تولید پایگاه دانش بهره برد. این سیستم ها دو هدف عمده را دنبال می­کنند:

* سازمان دهی اطلاعات داخل متن به نحوی کارا برای استفاده.
* قرار دادن اطلاعات در یک فرم دقیق منطقی که اجازه استنتاج و بررسی بیشتر را به الگوریتم ها و روش های کامپیوتری دیگر می­دهد.

بسیاری از الگوریتم های پیشرفته پردازش متن نمی­توانند با متن ورودی به صورت خام کار کنند و نیازمند دریافت برخی از اطلاعات به صورت ساختیافته هستند. مثلا به برچسب های مقوله دستوری[[3]](#footnote-3) یا اسامی موجود در متن یا نوع ارتباط آنها با هم در متن، نیازمندند.

سیستم های استخراج دانش معمولا دانشی که به صورت حقایق آشکار[[4]](#footnote-4) است را از متن بیرون می­کشند. اطلاعاتی نظیر «چه کسی در چه زمانی چه کاری را با چه کسی انجام داد؟». به عنوان نمونه درآمد یک نفر یا شرکت، دارایی های او، اعضای هیئت مدیره، دفاتر یک شرکت، دفتر مرکزی، مدیر عامل و استخراج هرگونه حقایق از این دست که از درون گزارش ها می­توان بیرون کشید. یا اسخراج کتاب های یک نویسنده، کارگردان یک فیلم سینمایی، محصولات یک کمپانی تولید خودرو یا نام مخترع یک وسیله با پردازش متون موجود در اینترنت. به عنوان نمونه به جمله زیر توجه کنید.

*اداره کل آموزش و پرورش و مراکز دفتری استانهای آن، در خیابان طالقانی تهران قرار دارند.*

ایده این است که سیستم بتواند جمله بالا را پردازش کرده و از داخل آن متوجه شود که «اداره کل آموزش و پرورش» و «مراکز دفتری استانها» نام دو سازمان است و به آنها برچسب ORG بدهد همچنین متوجه شود که این جمله درباره مراکز دفتری صحبت میکند و محل قرار گیری آنها را تعیین نموده است. بنابراین خروجی زیر را تولید کند:

Located(«اداره کل آموزش و پرورش», «خیابان طالقانی تهران»)

Located(«مراکز دفتری استانهای آموزش و پرورش», «خیابان طالقانی تهران»)

استخراج دانش همچنین در یافتن اطلاعات از متون پزشکی-زیست شناسی مانند روابط بین دارو های مختلف و بیماری ها، تاثیراتی ژنتیکی داروهای مختلف، تاثیر ژن در بیماری ها، اسامی گونه های مختلف گیاهان و ... مورد استفاده فراوانی قرار گرفته است.

اگر با دید کاربردی به استخراج دانش نگاه کنیم؛ این علم هم اکنون در بسیاری از تکنولوژی های امروزی موجود بوده و حتی فراگیر شده است. در بسیاری از کاربردها مانند سرویس ایمیل گوگل یا اپل یا اندیس گذاری اینترنتی. همانطور که در تصویر زیر مشاهده میکنید، سرویس ایمیل اپل به صورت اتوماتیک تاریخ و ساعت هایی که درون متن باشد را تشخیص می­دهد و این امکان را به کاربر می­دهد که تنها با یک کلیک آن را در تقویم خود برای یادآوری ذخیره کند.



مثالی دیگر در جستجوی گوگل قابل مشاهده است. اگر به دنبال مقر اصلی یک شرکت چند ملیتی بگردیم؛ گوگل همراه با بازگرداندن نتایج جستجو، بهترین حدس خود را برای پاسخ به سوال ما، نیز در ابتدا لیست نمایش می­دهد. به تصویر زیر توجه کنید.



یک زیر کلاس بسیار مهم از استخراج دانش، شناسایی موجودیت نامدار است که در آن، به دنبال یافتن و دسته بندی موجودیت های نامدار (اسامی) موجود در متن هستیم. بنابراین قدم اول یافتن اسامی است. در واقع محدوده اسامی در متن را تعیین میکنیم.

*سرانجام در سومین روز هفدهمین دوره مسابقات آسیایی در اینچئون کره‌جنوبی، طلسم بی‌مدالی کاروان ایران شکسته شد و نجمه خدمتی، تیرانداز هجده‌ساله تیم ملی کشورمان با درخشش خیره‌کننده‌ای به مقام قهرمانی آسیا رسید و نخستین طلای این رقابت‌ها را به نام کشورمان سند زد تا فراموش کنیم، ‌تیم فوتبال امیدکشورمان امروز‌ رسما از رقابت‌های آسیایی اینچئون حذف شد*

قدم بعدی دسته بندی اسامی یافته شده است.

*سرانجام در سومین روز هفدهمین دوره مسابقات آسیایی در اینچئون کره‌جنوبی، طلسم بی‌مدالی کاروان ایران شکسته شد و نجمه خدمتی، تیرانداز هجده‌ساله تیم ملی کشورمان با درخشش خیره‌کننده‌ای به مقام قهرمانی آسیا رسید و نخستین طلای این رقابت‌ها را به نام کشورمان سند زد تا فراموش کنیم، ‌تیم فوتبال امیدکشورمان امروز‌ رسما از رقابت‌های آسیایی اینچئون حذف شد*

قرمز: سازمان – سبز: شهر – آبی: کشور – نارنجی: نام فرد – بنفش: قاره – سبز تیره: تاریخ

موجودیت[[5]](#footnote-5) به چیزی خیالی یا واقعی گفته می‌شود که به ذات خویش وجود دارد. به عنوان مثال هر انسان یک موجودیت است، دانشگاه شهید بهشتی، ایران یا شرکت مخابرات ایران یک موجودیت هستند. اما هوا یا خاک موجودیت نیستند. ایده پشت موجودیت نامدار این است که برخی از موجودیت ها دارای نام هستند. مثلا من محمود راحت نام دارم، کشورم ایران نام دارد، شهرم تهران نام دارد اما صندلی که روی آن نشسته ام با این حال که یک موجودیت است؛ اما نام ندارد. البته بحث اینکه چه چیزهایی موجودیت نامدار هستند و چه چیزهایی نستند، کمی فلسفی است و دیدگاه های مختلفی در این زمینه وجود دارد. اما در این بحث ما به دنبال اسامی و موجودیت های نامداری هستیم که به راحتی قابل تمیز باشند و از نظر فلسفی رو موجودیت بودن آنها شکی نیست. همچنین در پردازش متن چیزهایی مثل تاریخ (یازده خرداد سال 1364) و زمان (ساعت ده شب – 11:30) و مقدار و نام مواد شیمیایی، پروتئنی یا زیستی نیز به عنوان موجودیت های نامدار در نظر گرفته می­شوند. با این حال که آنها به هیچ عنوان موجودیت نامدار نیستند ولی در زمان تعیین اسامی نامدار در متن عموما این اسامی نیز در نظر گرفته می­شوند.

در زیر برخی از کاربرد های شناسایی موجودیت های نامدار را مشاهده می­کنید:

* از موجودیت های نامدار در متن می­توان برای اندیس گذاری یا لینک دادن استفاده کرد. (بسیاری از شرکت ها از روشهای مختلف برای استخراج اسامی در متن استفاده میکنند و در صفحات وب خود به آنها لینک می­زنند. مثلا در صفحات وب بایولوژی نام پروتئین ها و ژن های متن به صورت لینک بوده و به صفحه ای که به تبیین آن موجودیت می­پردازد لینک شده اند.)
* در کاربرد تشخیص نظر[[6]](#footnote-6) کاربران، نیازمندیم که بدانیم نظر نویسنده متن در مورد یک موضوع خاص مثبت یا منفی است. علاوه بر آن نیازمندیم بدانیم این نظر در مورد چه کسی یا چه چیزی است. لذا استخراج موجودیت های نامدار (شامل نام شرکت و نام محصول) می­تواند به تشخیص درست نظر کاربران کمک نماید.
* اگر به شناسایی موجودیت های نامدار به عنوان پیش نیاز استخراج دانش نگاه کنیم، کاربردهای زیادی برای آن وجود دارد. در استخراج دانش بیشتر کاری که انجام می­گیرد، استخراج موجودیت های نامدار و سپس تعیین روابط بین آنها است.
* کاربرد در سیستم های پرسش و پاسخ. در این سیستم ها پاسخ بسیاری از انواع سوالات، موجودیت های نامدارند. اینکه یک چیز چه زمانی اتفاق افتاده یا چه کسی آن را انجام داده است. بنابراین اگر یک سیستم خوب استخراج موجودیت نامدار برای زبان داشته باشیم، کمک شایانی به پاسخ سوالات به صورت اتوماتیک می­کند.

در ادامه برخی از سایت ها و پروژه های موجود در زمینه استخراج دانش برای زبان انگلیسی معرفی می­شوند.

پروژه Open Calais از خبرگزاری Reuter’s یک نمونه جالب از استخراج اطلاعات است. در این سایت بخشی به شکل دموی آنلاین وجود دارد که میتوان یک متن انگلیسی دلخواه را در آن قرار داد. در شکل زیر نمونه خروجی آن را مشاهده میکنید.







همچنین با قرار دادن کرسر موس بر روی هر کدام از موجودیت های نامدار دانش استخراج شده از متن برای آن نمایش داده میشود. شکل زیر.



پروژه AlchemyAPI برای 8 زبان انگلیسی، فرانسه، آلمانی، ایتالیایی، پرتغالی، روسی، اسپانیایی و سوئدی قابلیت استخراج موجودیت های نامدار را دارد. در شکل زیر نمونه ای از خروجی این سیستم را مشاهده میکنید. متن ورودی، همان متن وارد شده در مثال بالا است.



از دیگر وب سایت ها میتوان به Evri و Yahoo’s term extraction اشاره نمود.

**ارزیابی سیستم های تشخیص موجودیت نامدار**

در این سیستمها، ورودی دنباله ای از کلمات و هدف شناسایی موجودیت نامدار است. برای اینکار باید شروع و پایان هر موجودیت و دسته مربوط به آن در متن تعیین گردد. به مثال زیر توجه کنید

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *رامین* | *مهمان* | *پرست،* | *سخنگوی* | *وزارت* | *امور* | *خارجه* | *دیروز* | *در* | *جمع* | *خبرنگاران* | *گفت* |
| PER | PER | PER | O | ORG | ORG | ORG | O | O | O | O | O |

به این مسئله می­توان به صورت یک مسئله دسته بندی کننده نگاه کرد که به هر کلمه ورودی یک برچسب می­زند. اما اینکار چندان درست نیست. زیرا در این مسئله ورودی دنباله ای از کلمات است و مثلا اگر یک کلمه با برچسب ORG داشته باشیم، احتمال اینکه کلمه بعدی نیز ORG باشد زیاد است. بنابراین با روش ذکر شده همبستگی بین توکن ها در نظر گرفت نشده است.

استفاده از روش معمول دقت و فراخوان[[7]](#footnote-7) برای شناسایی موجودیت های نامدار چندان کارا نیست. به مثال زیر توجه کنید.

|  |  |  |  |
| --- | --- | --- | --- |
| *رامین* | *مهمان* | *پرست،* | *سخنگوی* |
| PER | PER | O | O |

در این مثال برچسب زننده ما بدنبال یافتن اسامی در متن است. به درستی دو توکن اول را برچسب PER زده است ولی توکن سوم را به اشتباه برچسب O زده است. در روش دقت و فراخوان چون کل عبارت «رامین مهمان پرست» یک نام است؛ اگر سیستم کل عبارات را درست تشخیص دهد، یکی به تعداد درست های تشخیص داده شده اضافه می­شود و در غیر این صورت یکی به تعداد غلط ها. حال در این شرایط علارغم اینکه سیستم جوابی نزدیک به درست داده و تنها بخشی از پاسخ را جا انداخته است؛ دو امتیاز منفی یکی برای اینکه پاسخ اش درست نبوده و دیگری برای اینکه یک نام را در متن نیافته است می­گیرد. این در حالیست که اگر سیستم کلا به همه توکن ها برچسب O میزد تنها یک خطا داشت. پر واضح است که این خلاف منطق است.

برای حل این مشکل پیشنهادات مختلفی برای ارزیابی سیستم های شناسایی موجودیت نامدار ارائه شده است. در این سیستم ها به برچسب زن، اعتبار جزئی برای حالاتی که موجودیت نامدار را به شکل تقریبا صحیح تشخیص داده است (مانند مثال بالا) می­دهند. یکی از این روشها امتیاز ماک[[8]](#footnote-8) است. البته چون محاسبه خود امتیاز ماک هم پیچیدگی های خاص خود را دارد؛ زیرا باید تعیین کنیم چگونه و به هر پاسخ نسبی چه امتیازی داده شود؛ بنابراین هنوز هم در بسیاری از منابع و مقالات از معیار اف یک[[9]](#footnote-9) برای گزارش نتایج سیستم خود استفاده می­کنند.

1. Information extraction (IE) [↑](#footnote-ref-1)
2. Named entity recognition (NER) [↑](#footnote-ref-2)
3. Part of speech tags (POS) [↑](#footnote-ref-3)
4. Clear factual information [↑](#footnote-ref-4)
5. Entity [↑](#footnote-ref-5)
6. Sentiment analysis (opinion mining) [↑](#footnote-ref-6)
7. Precision recall [↑](#footnote-ref-7)
8. MUC score [↑](#footnote-ref-8)
9. F1 measure [↑](#footnote-ref-9)